CLAIMS

What is claimed is:

1. A method of evaluating a document set commonality for a document set including a plurality of documents each having one or more document segments, the commonality indicating a degree to which topics of the individual documents of the document set are common, the method comprising:

(a) generating for each of the document segments, a document segment vector in which components corresponding to terms appearing in the document segment assume a value "1" (one), and the other components assume a value "0" (zero);

(b) generating a co-occurrence matrix from the document segment vectors for each of the documents of the document set;

(c) generating a common co-occurrence matrix having rows and columns in which products of values of components of the same rows and the same columns of such co-occurrence matrices of the respective documents are given as values of components of the rows and the columns; and

(d) evaluating the document set commonality on the basis of a sum of either all the components or diagonal components of the common co-occurrence matrix.

2. A method of evaluating either a document - document set commonality or a document segment - document set commonality for a document set including a plurality of documents each having one or more document segments, the commonality indicating a degree to which each of the documents or each of the document segments is close to a topic common to the document set, the method comprising:

(a) generating for each of the document segments, a document segment vector in which components corresponding to terms appearing in the document segment assume a value "1" (one), and the other components assume a value "0" (zero);

(b) generating a co-occurrence matrix from the document segment vectors for each of the documents of the document set;

(c) generating a common co-occurrence matrix having rows and columns in which products of values of components of the same rows and the same columns of such co-occurrence matrices of the respective documents are given as values of components of the rows and the columns; and

(d) evaluating either the document - document set commonality or the document segment - document set commonality on the basis of a product-sum between either the co-occurrence matrices of the documents or the document segments and either all the components or diagonal components of the common co-occurrence matrix.

3. A method of calculating a document set commonality for a document set including a plurality of documents each having one or more document segments, the method comprising:

(a) generating for each of the document segments, a document segment vector in which components corresponding to terms appearing in the document segment assume a value "1" (one), and the other components assume a value "0" (zero);

(b) generating a co-occurrence matrix from the document segment vectors for each of the documents of the document set;

(c) generating a common co-occurrence matrix having rows and columns of a mismatch allowance type from products of values of components of the same rows and the same columns of such co-occurrence matrices of the respective documents, except in cases where the values of the components of the same rows and the same columns are "0" (zero);

(d) checking if the components of the co-occurrence matrices of the respective documents have the value "0", and creating a co-occurrence count matrix for counting

the number of the documents whose components are not "0"; and

(e) correcting when each of components of the co-occurrence count matrix has a value less than a predetermined threshold, the corresponding component of the common co-occurrence matrix of mismatch allowance type so as to become "0", and evaluating the document set commonality of mismatch allowance type on the basis of a sum of either all the components or diagonal components of the corrected common co-occurrence matrix of a mismatch allowance type.

4. A method of extracting documents of common topic from within a document set including a plurality of documents each having one or more document segments; the method comprising:

(a) generating for each of the document segments, a document segment vector in which components corresponding to terms appearing in the document segment assume a value "1" (one), and the other components assume a value "0" (zero);

(b) generating a co-occurrence matrix from such document segment vectors for each of the documents of the document set;

(c) generating a common co-occurrence matrix having rows and columns of a mismatch allowance type from products of values of components of the same rows and the same columns of such co-occurrence matrices of the respective documents, except in cases where the values of the components of the same rows and the same columns are "0" (zero);

(d) checking if the components of the co-occurrence matrices of the respective documents have the value "0", and creating a co-occurrence count matrix for counting the number of the documents whose components are not "0";

(e) correcting when each of components of the co-occurrence count matrix has a value less than a predetermined threshold, the corresponding component of the common

co-occurrence matrix of a mismatch allowance type so as to become "0", and evaluating a document set commonality of a mismatch allowance type on the basis of a sum of either all the components or diagonal components of the corrected common co-occurrence matrix of mismatch allowance type;

(f) evaluating a mismatch allowance type document - document set common commonality when the document set commonality of a mismatch allowance type is not less than a certain threshold, for each of the documents and on the basis of either a product-sum between all the components of the co-occurrence matrix of the document and all the components of the corrected common co-occurrence matrix of mismatch allowance type or a product-sum between the diagonal components of the co-occurrence matrix of the document and the diagonal components of the corrected common co-occurrence matrix of a mismatch allowance type; and

(g) extracting the documents as to which the mismatch allowance type document - document set common commonality exceeds a predetermined threshold, as the documents of a common topic.

5. The method as defined in claim 1, further comprising letting M denote the number of sorts of the occurring terms, D_r denote an r*th* document in a document set D consisting of R documents, Y_r denote the number of document segments of the document D_r, and $d_{ry} = (d_{ry1}, \ldots, d_{ryM})^T$ denote a y*th* document segment vector of the document D_r, letter T indicating transposition of a vector, and determining the co-occurrence matrix S^r of the document D_r by:

$$S^r = \sum_{y=1}^{Y_r} d_{ry} d_{ry}^T$$

6. The method as defined in claim 2, further comprising letting M denote the number of sorts of the occurring terms, D_r denote an r*th* document in a document set D consisting of R documents, Y_r denote the number of document segments of the document D_r, and $d_{ry} = (d_{ry1}, \ldots, d_{ryM})^T$ denote a y*th* document segment vector of the document D_r, letter T indicating transposition of a vector, and determining the co-occurrence matrix S^r of the document D_r by:

$$S^r = \sum_{y=1}^{Y_r} d_{ry} d_{ry}^T$$

7. The method as defined in claim 3, further comprising letting M denote the number of sorts of the occurring terms, D_r denote an r*th* document in a document set D consisting of R documents, Y_r denote the number of document segments of the document D_r, and $d_{ry} = (d_{ry1}, \ldots, d_{ryM})^T$ denote a y*th* document segment vector of the document D_r, letter T indicating transposition of a vector, and determining the co-occurrence matrix S^r of the document D_r by:

$$S^r = \sum_{y=1}^{Y_r} d_{ry} d_{ry}^T$$

8. The method as defined in claim 4, further comprising letting M denote the number of sorts of the occurring terms, D_r denote an r*th* document in a document set D consisting of R documents, Y_r denote the number of document segments of the document D_r, and $d_{ry} = (d_{ry1}, \ldots, d_{ryM})^T$ denote a y*th* document segment vector of the document D_r, letter T indicating transposition of a vector, and determining the co-occurrence matrix S^r of the document D_r by:

$$S^r = \sum_{y=1}^{Y_r} d_{ry} d_{ry}^T$$

9. The method as defined in claim 1, further comprising determining an mn component S^C_{mn} of a common co-occurrence matrix S^C of a document set D by:

$$S^C_{mn} = \prod_{r=1}^{R} S^r_{mn}$$

10. The method as defined in claim 2, further comprising determining an mn component S^C_{mn} of a common co-occurrence matrix S^C of a document set D by:

$$S^C_{mn} = \prod_{r=1}^{R} S^r_{mn}$$

11. The method as defined in claim 3, further comprising determining an mn component S^C_{mn} of a common co-occurrence matrix S^C of a document set D by:

$$S^C_{mn} = \prod_{r=1}^{R} S^r_{mn}$$

12. The method as defined in claim 4, further comprising determining an mn component S^C_{mn} of a common co-occurrence matrix S^C of a document set D by:

$$S^C_{mn} = \prod_{r=1}^{R} S^r_{mn}$$

13. The method as defined in claim 1, further comprising each diagonal component of a common co-occurrence matrix S^C of a document set D being approximated by a product of occurring frequencies of the corresponding term in the respective documents.

14. The method as defined in claim 2, further comprising each diagonal component of a common co-occurrence matrix S^C of a document set D being approximated by a product of occurring frequencies of the corresponding term in the respective documents.

15. The method as defined in claim 3, further comprising each diagonal component of a common co-occurrence matrix S^C of a document set D being approximated by a product of occurring frequencies of the corresponding term in the respective documents.

16. The method as defined in claim 4, further comprising each diagonal component of a common co-occurrence matrix S^C of a document set D being approximated by a product of occurring frequencies of the corresponding term in the respective documents.

17. A program storage device, readable by a machine, tangibly embodying a program of instructions executable by the machine to perform the method of claim 1.

18. A program storage device, readable by a machine, tangibly embodying a program of instructions executable by the machine to perform the method of claim 2.

19. A program storage device, readable by a machine, tangibly embodying a program of instructions executable by the machine to perform the method of claim 3.

20. A program storage device, readable by a machine, tangibly embodying a program of instructions executable by the machine to perform the method of claim 4.

21. The program storage device as defined in claim 17, further comprising letting M denote the number of sorts of the occurring terms, D_r denote an r*th* document in a document set D consisting of R documents, Y_r denote the number of document segments of the document D_r, and $d_{ry} = (d_{ry1}, \ldots, d_{ryM})^T$ denote a y*th* document segment vector of the document D_r, letter T indicating transposition of a vector, determining the co-occurrence matrix S^r of the document D_r by:

$$S^r = \sum_{y=1}^{Y_r} d_{ry} d_{ry}^T$$

22. The program storage device as defined in claim 18, further comprising letting M denote the number of sorts of the occurring terms, D_r denote an r*th* document in a document set D consisting of R documents, Y_r denote the number of document segments of the document D_r, and $d_{ry} = (d_{ry1}, \ldots, d_{ryM})^T$ denote a y*th* document segment vector of the document D_r, letter T indicating transposition of a vector, determining the co-occurrence matrix S^r of the document D_r by:

$$S^r = \sum_{y=1}^{Y_r} d_{ry} d_{ry}^T$$

23. The program storage device as defined in claim 19, further comprising letting M denote the number of sorts of the occurring terms, D_r denote an r*th* document in a document set D consisting of R documents, Y_r denote the number of document segments of the document D_r, and $d_{ry} = (d_{ry1}, \ldots, d_{ryM})^T$ denote a y*th* document segment vector of

the document D_r, letter T indicating transposition of a vector, determining the co-occurrence matrix S^r of the document D_r by:

$$S^r = \sum_{y=1}^{Y_r} d_{ry} d_{ry}^T$$

24. The program storage device as defined in claim 20, further comprising letting M denote the number of sorts of the occurring terms, D_r denote an r*th* document in a document set D consisting of R documents, Y_r denote the number of document segments of the document D_r, and $d_{ry} = (d_{ry1}, \ldots, d_{ryM})^T$ denote a y*th* document segment vector of the document D_r, letter T indicating transposition of a vector, determining the co-occurrence matrix S^r of the document D_r by:

$$S^r = \sum_{y=1}^{Y_r} d_{ry} d_{ry}^T$$

25. The program storage device as defined in claim 17, further comprising determining an mn component $S^C{}_{mn}$ of a common co-occurrence matrix S^C of a document set D by:

$$S^C{}_{mn} = \prod_{r=1}^{R} S^r{}_{mn}$$

26. The program storage device as defined in claim 18, further comprising determining an mn component $S^C{}_{mn}$ of a common co-occurrence matrix S^C of a document set D by:

$$S^C{}_{mn} = \prod_{r=1}^{R} S^r{}_{mn}$$

27. The program storage device as defined in claim 19, further comprising determining an mn component S^C_{mn} of a common co-occurrence matrix S^C of a document set D by:

$$S^C_{mn} = \prod_{r=1}^{R} S^r_{mn}$$

28. The program storage device as defined in claim 20, further comprising determining an mn component S^C_{mn} of a common co-occurrence matrix S^C of a document set D by:

$$S^C_{mn} = \prod_{r=1}^{R} S^r_{mn}$$

29. A computer system arranged to perform the method of claim 1.

30. A computer system arranged to perform the method of claim 2.

31. A computer system arranged to perform the method of claim 3.

32. A computer system arranged to perform the method of claim 4.